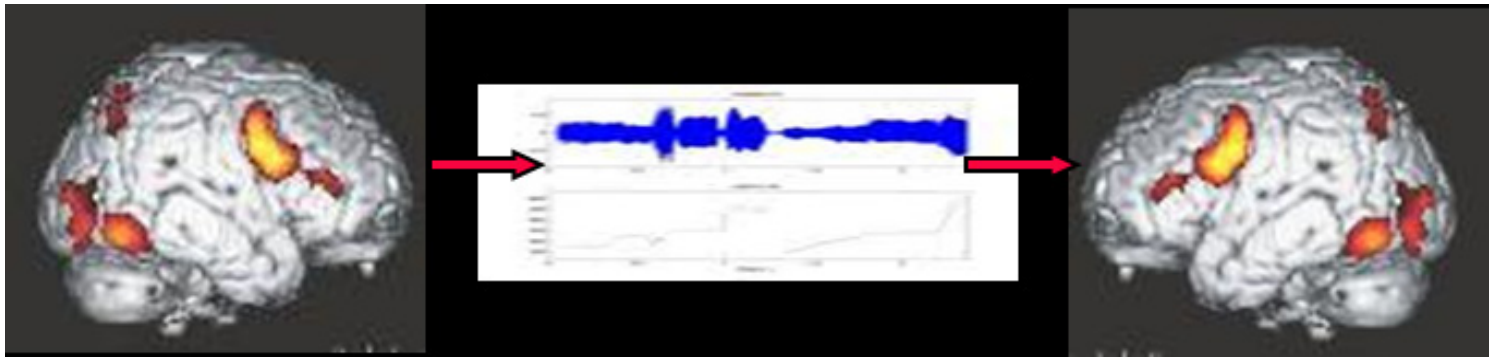
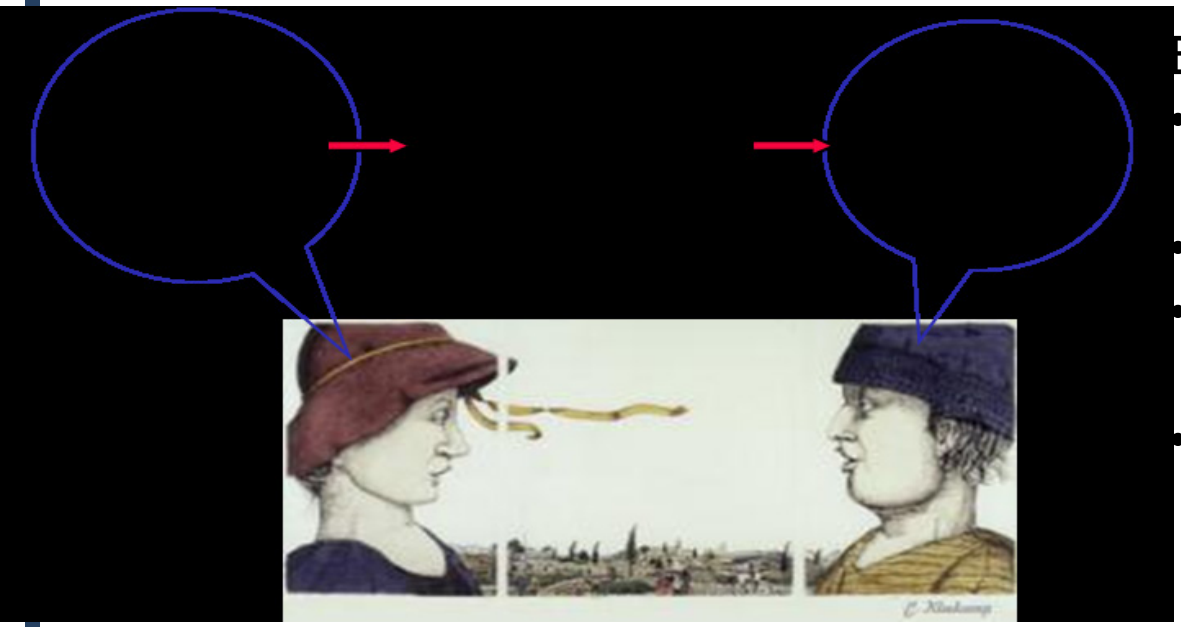
Wege zu globalen Daten-Infrastrukturen

Peter Wittenburg
Max Planck Data and Compute Center
Garching, Deutschland
Max Planck Institute for Psycholinguistics
Nijmegen, Niederlande

mein Hintergrund

Psycholinguistik



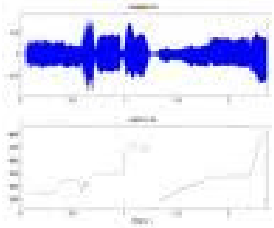
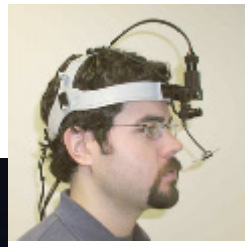
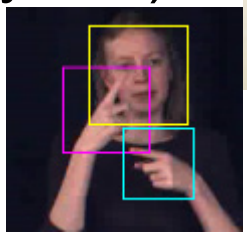


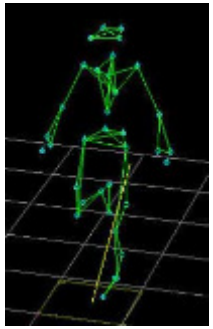
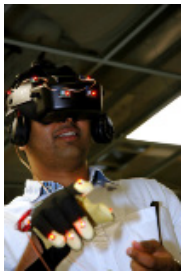
Experimente

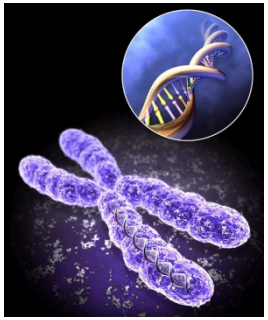
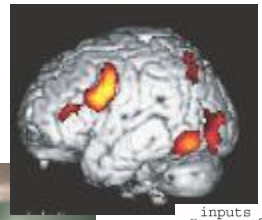
- wie verarbeitet das Gehirn Sprache?
- wie erlernen wir Sprache?
- wie ist die funktionelle Architektur?
- was ist genetisch vorbestimmt?

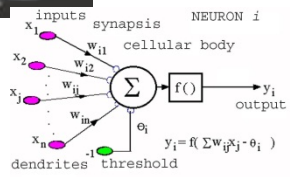
Experimente & Observationen

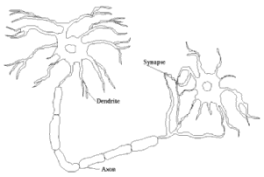
- nutzen alle verfügbaren Kanäle
  - speech sounds
  - suprasegmental information (pitch, intensity, etc)
  - eye movements
  - head movements
  - hand/arm movements (gestures)
  - body movements
  - virtual reality
  - EEG/MEG/fMRI
  - genomics
  - simulations
  - etc.



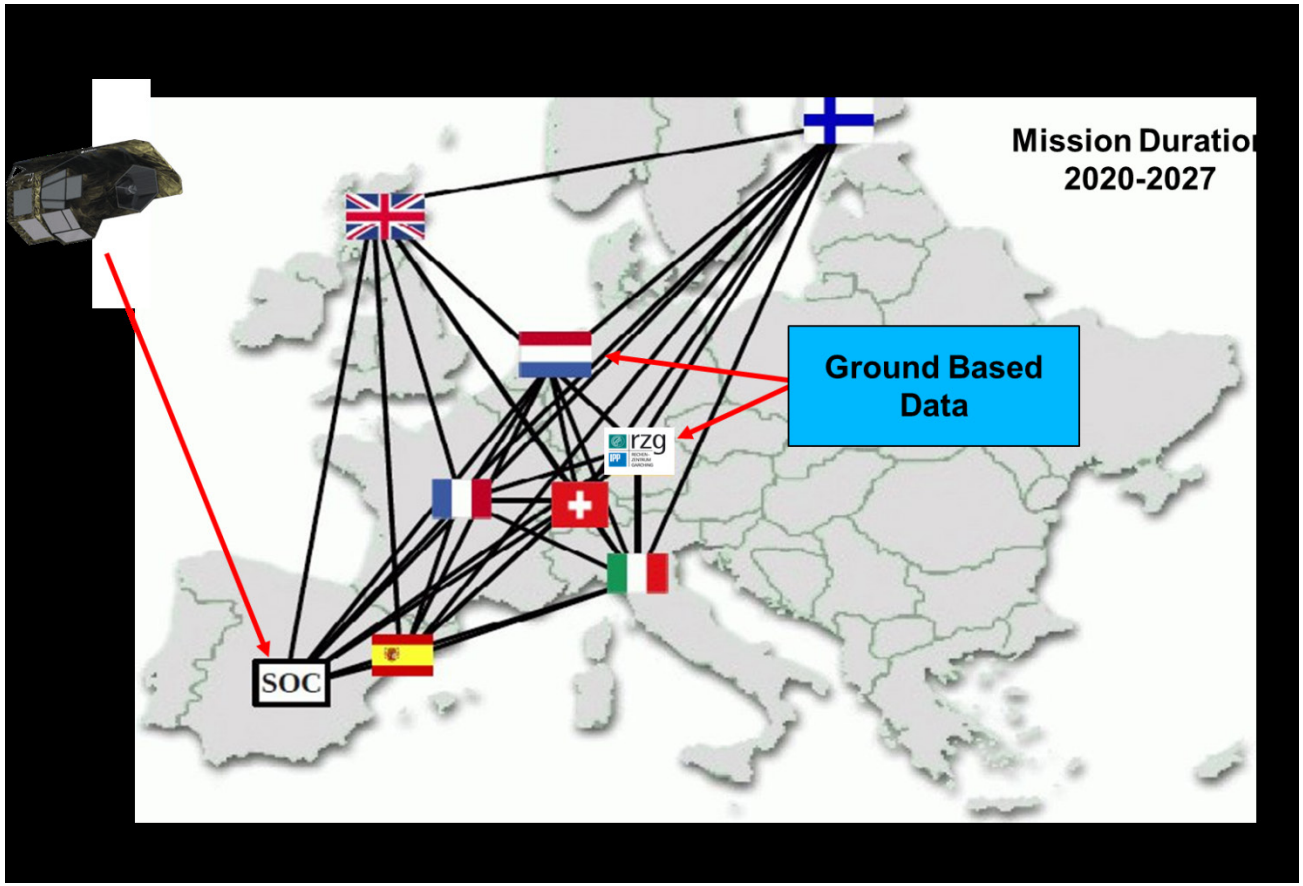


Wissenschaft unter dem Einfluss der Daten

einige MPIs als Beispiel

- GW-Sektion
- BM-Sektion
- PCT-Sektion

Globales EUCLID Projekt



- **1.0 PB raw data** from ground based surveys
- **300 TB Euclid** raw data
- processed data **5x more**

slide von Johannes Koppenhoefer, MPE

MPI for Neurobiologie

TECHNOLOGY FEATURE

CHARTING THE BRAIN'S NETWORKS

The field of connectomics is pulling neuroscience into a speedy, high-throughput lane that is generating vast amounts of data.

ALLEN INST. BRAIN SCI.

Massive stores of brain-tissue slides are providing a resource for scientists working on mapping neural networks.

BY VIVIEN MARX

neuroscientist Moritz Helmstaedter at the Max

web of around 100 billion neurons in the

- Electronen Microscopie neuronaler Strukturen
- Datenvolumen: **10 – 100PB**

"I think this is a really exciting field," says

power to map the massively interconnected

A few years ago, it was nearly impossible

DOBES Spracharchiv am MPI

- ca. 200 Sprachen und Kulturen – zumeist unwiderbringliche Aufnahmen
- ca. **80 TB** im online Repositorium (PIDs, MD)
- externe Replikationen über GWDG, RZG und evtl. SARA
- ca. **200 TB** nicht gut organisierte Daten
- DOBES Programm von der VWS seit 2000 gefördert

Multimodale Analysen am MPI

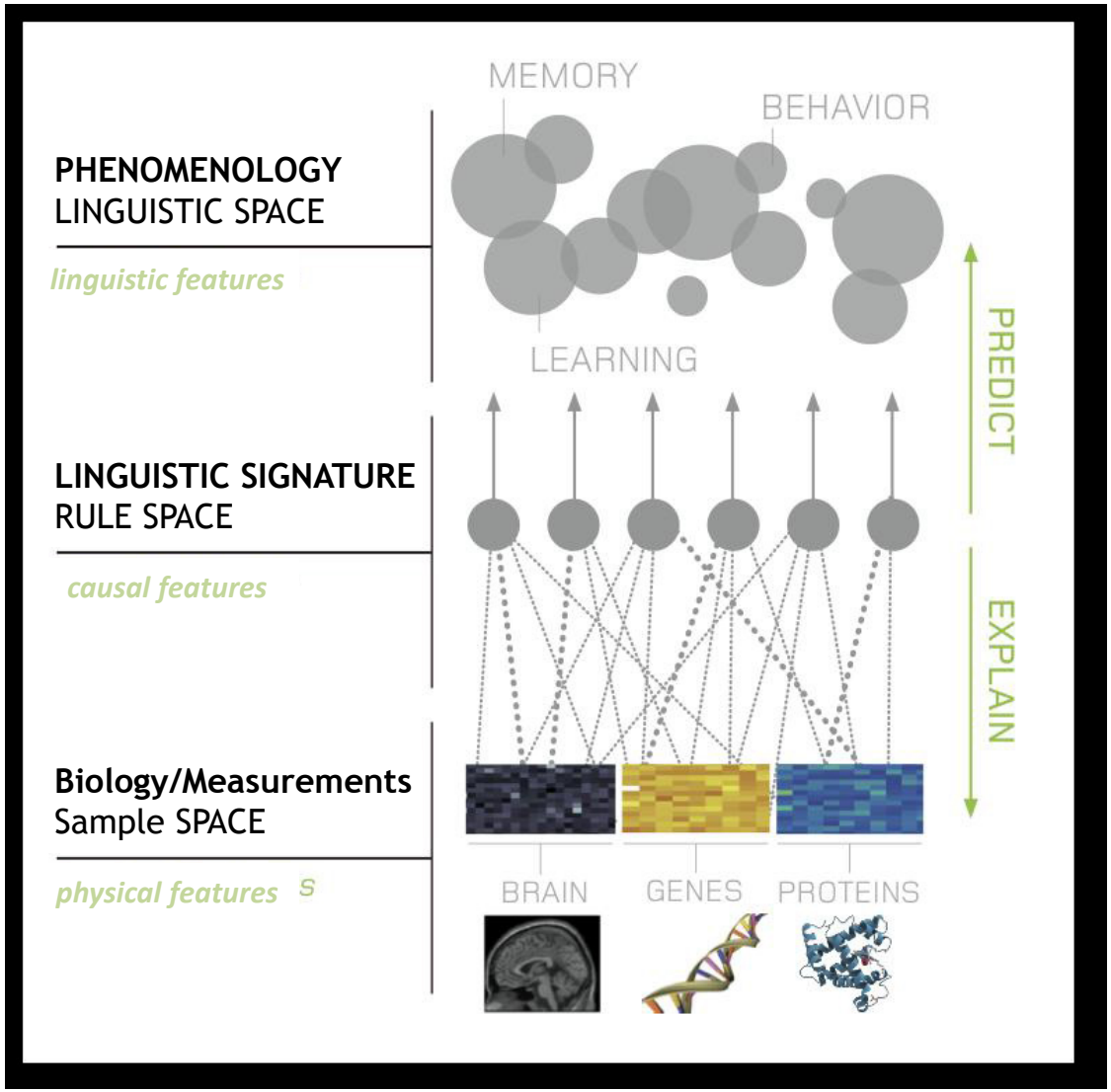
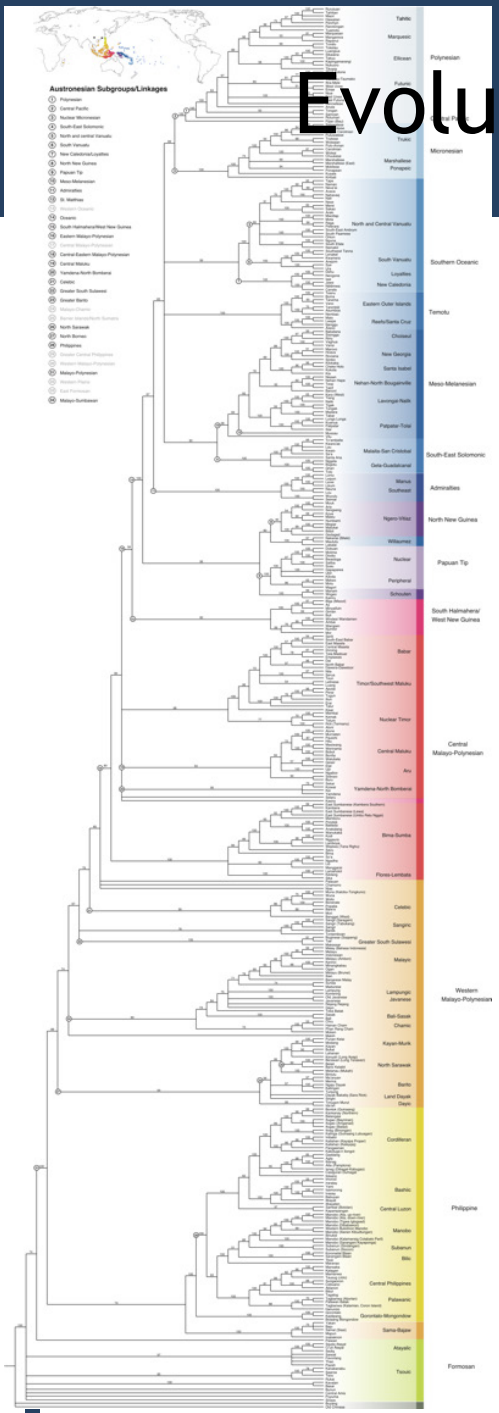


Bild vom CHUV & EPFL Lausanne

- Begründen linguistischer Phänomene mittels verschiedener Muster in unterschiedlichen Datenquellen (Resolution (T, SP), Art, VP, etc.)
- mittels ML eine Abbildung zwischen Samples auf Linguistische Phänomene
- ca. **2.5 TB Datenmatrix**
- Daten von diversen Instituten

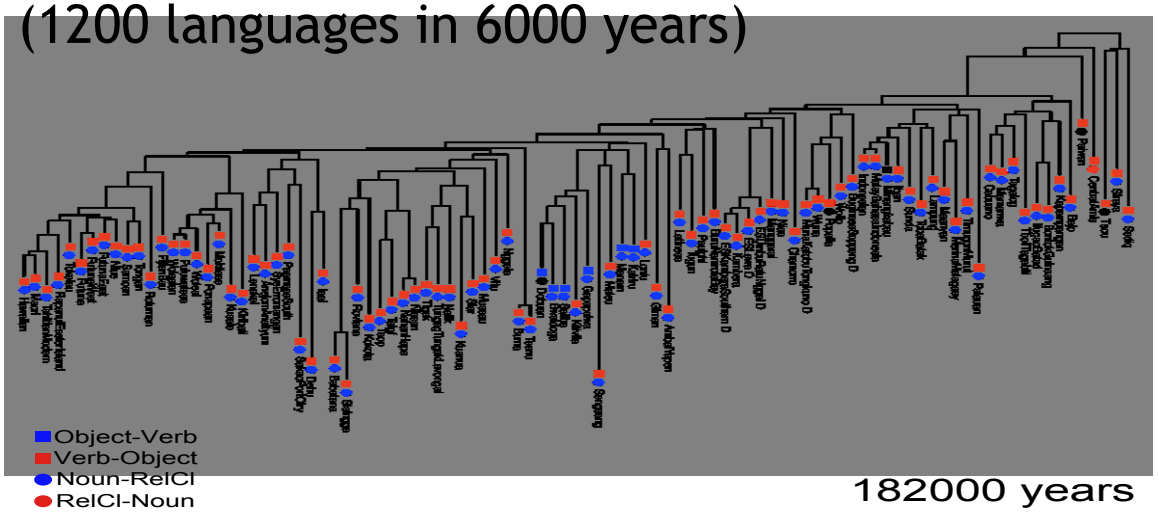
Evolution der Sprachen am MPI



was sind die Wurzeln unserer Sprachen?

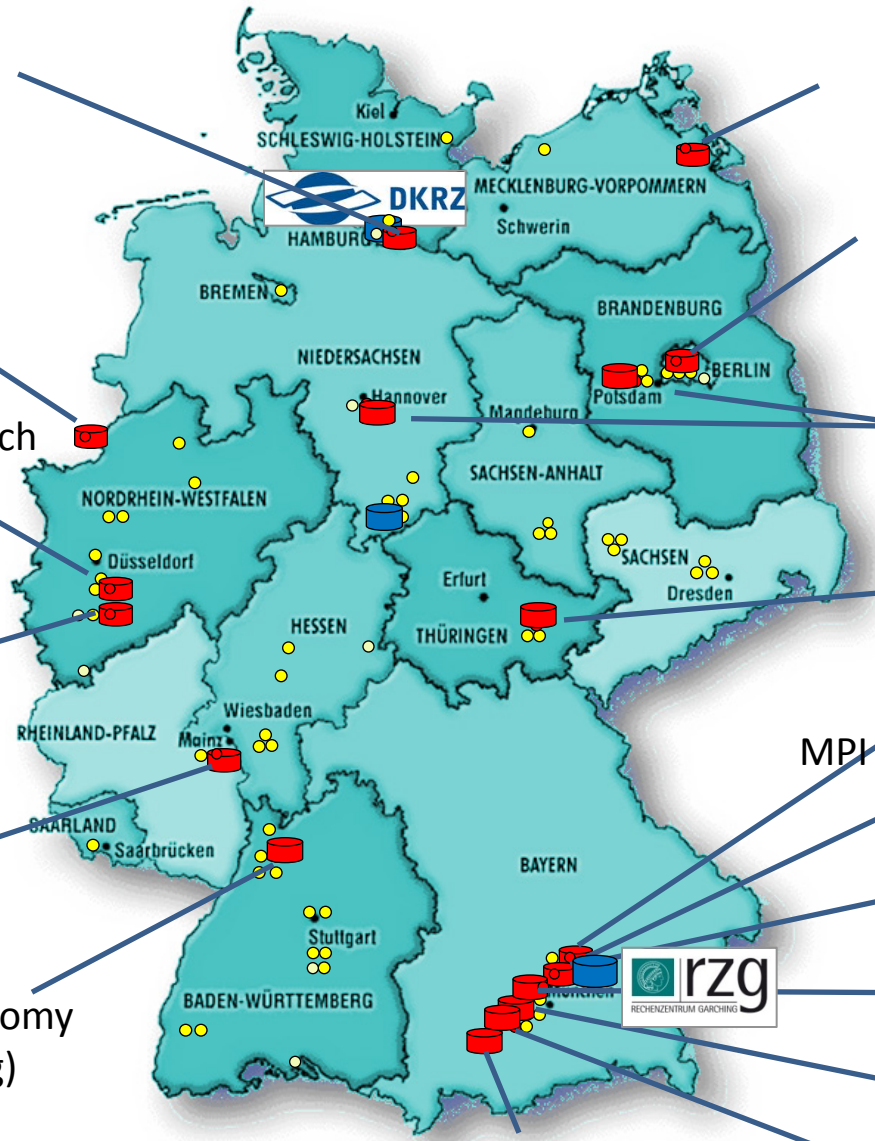
Austronesische Sprachen

- schier unglaubliche Proliferation der Diversität
- eine neue Sprache innerhalb von 5 Jahren (1200 languages in 6000 years)



- der Clou ist eine **große Featurematrix** und Erstellen von Entwicklungsbäumen mittels Phylogenetischer Algorithmen

MPIe mit großen Daten Projekten



MPI for Meteorology (Hamburg)

MPI for Psycholinguistics (Nijmegen)

MPI f. Plant Breeding Research (Köln)

MPI for Radio Astronomy (Bonn)

MPI for Chemistry (Mainz)

MPI for Astronomy (Heidelberg)

MPI for Plasmaphysics (Greifswald)

MPI for Molecular Genetics (Berlin)

MPI for Gravitational Physics (Potsdam, Hannover)

MPI for Biogeochemistry (Jena)

MPI for Plasmaphysics (Garching)

MPI for Extraterrestrial Physics (Garching)

MPI for Astrophysics (Garching)

MPI for Physics (München)

MPI for Neurobiology (Martinsried)

MPI for Biochemistry (Martinsried)

MPI for Ornitology (Seewiesen)

ein paar Anmerkungen

Daten-orientierte Wissenschaft ...

- generiert immer mehr Daten, die d … t zu einer Herausforderung werden l
  - reproduzierbare Wissenschaft
  - Vertrauen in Basis wissenschaftl
  - Steigern der Effizienz (50+ % Ve
  - gesellschaftliche Verantwortung
- ist kollaborativ, cross-diszipilinär u … grenzüberg…
- ist dynamisch im Erfinden neuer Str…
  semantischer Domänen
- braucht stabile und doch flexible Ra…

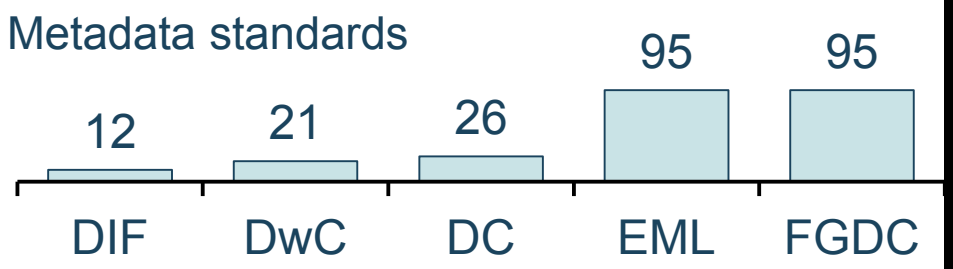
N. Kroes EC:

Data is currency of modern Science

Wille zum Austausch

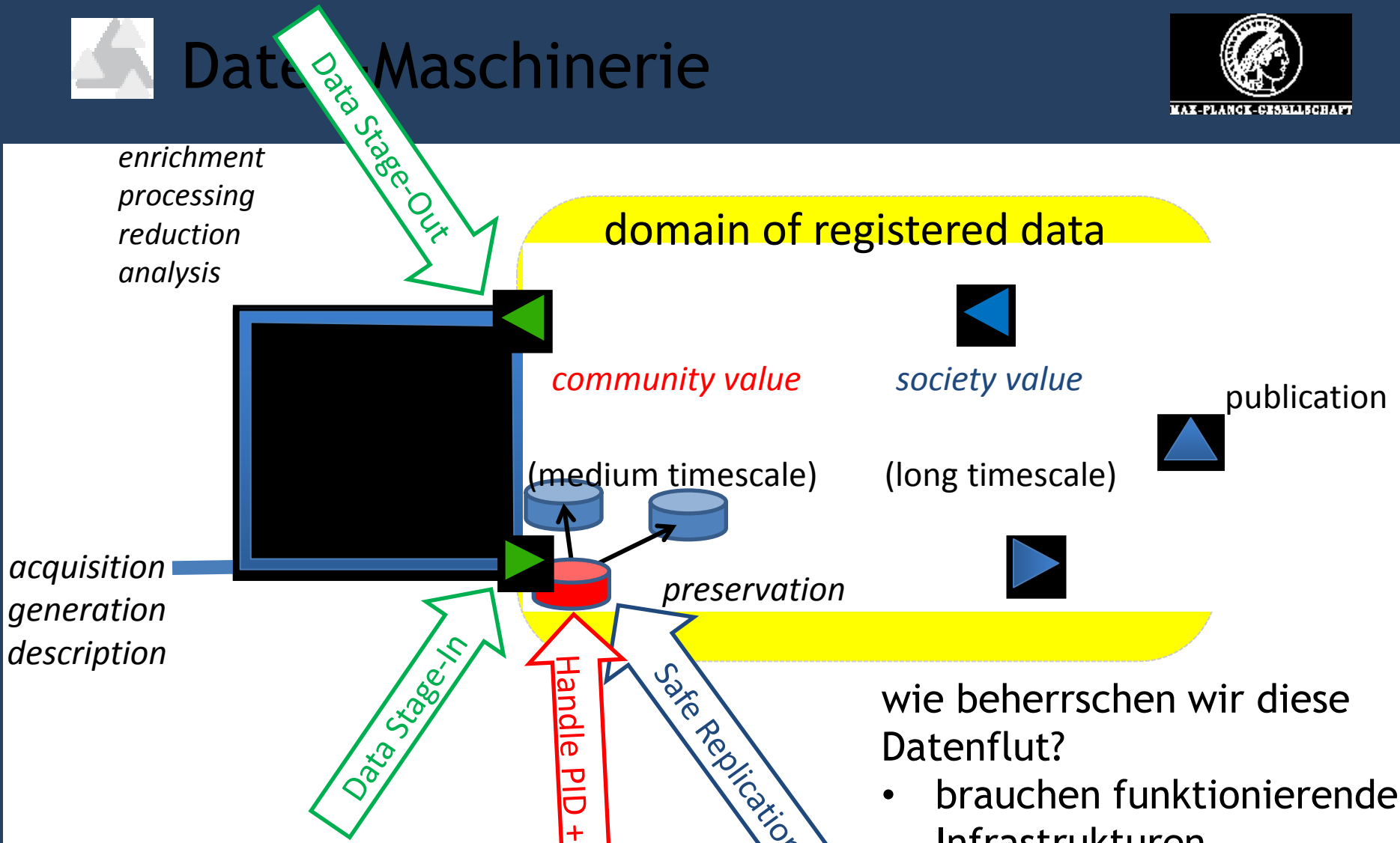
Use other researchers' datasets if easily accessible	84%
Willing to share data across a broad group of researchers	81%
Appropriate to create new datasets from shared data	76%
← Currently share all of their data	6%

…wie jedoch anstellen?



slide von Bill Michener, DataONE

Daten-Maschinerie



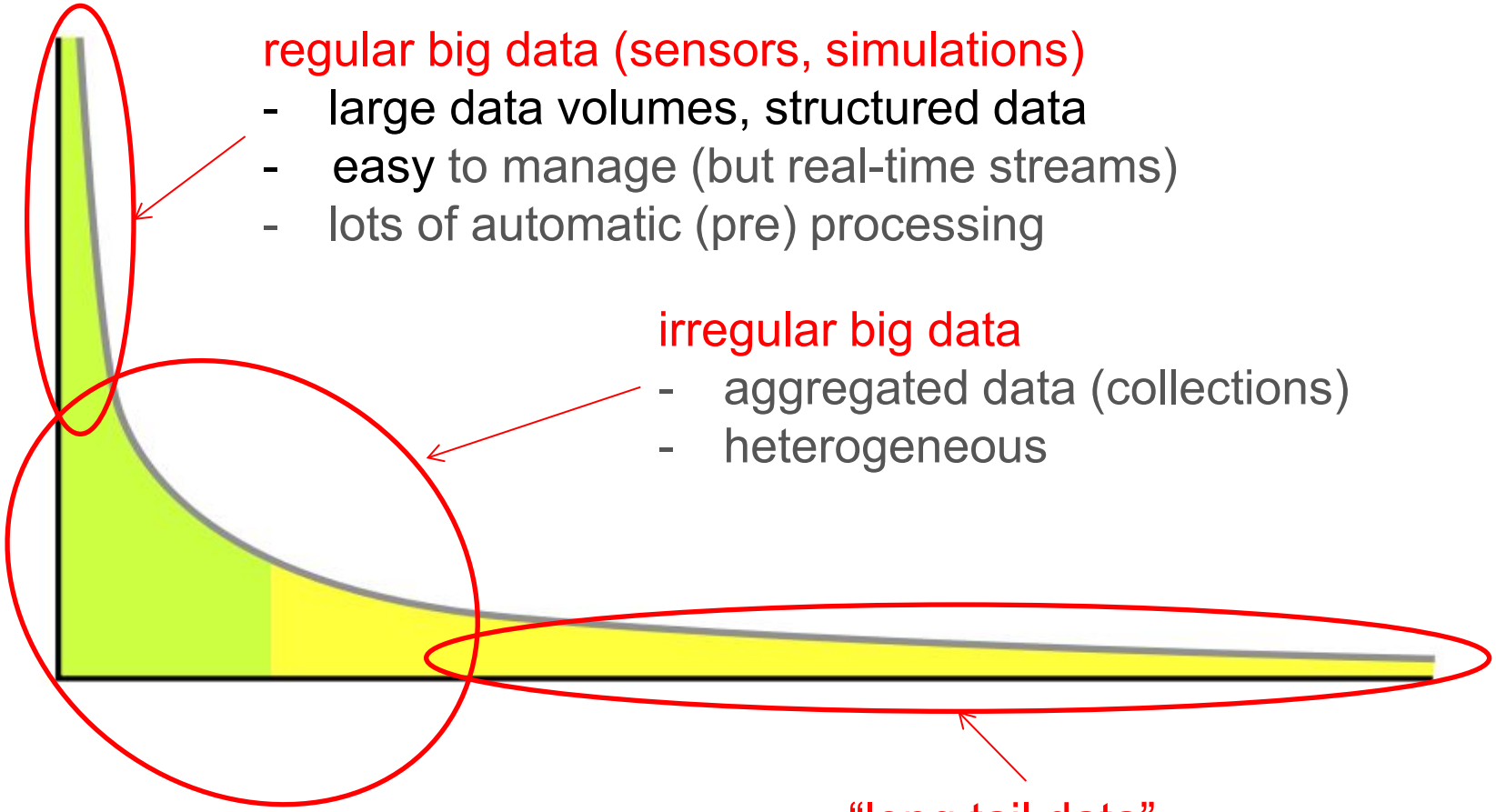
PIDs geben Identität und sichern Referenzierbarkeit MD erlauben Finden und Interpretation.

wie beherrschen wir diese Datenflut?

- brauchen funktionierende Infrastrukturen
- brauchen automatische Abläufe

Daten-Fabrik!

nicht nur Big Data



regular big data (sensors, simulations)
- large data volumes, structured data
- easy to manage (but real-time streams)
- lots of automatic (pre) processing

irregular big data
- aggregated data (collections)
- heterogeneous

“long tail data”
- difficult to manage
- difficult to get and harvest
- lots of relations

Wissenschaftler Workshop zu Date

- einige Daten-Challenges sind nur durch Wissenschaftler selbst zu lösen (Algorithmen,etc.) – andere bedürfen einer Infrastruktur
- nur systematischere Lösungen werden bei der Zunahme der Daten reproduzierbare Wissenschaft ermöglichen – Trend zu automatisierten Workflows
- Umgang mit Daten kostet insgesamt zu viel und belegt Wissenschaftler Zeit – “bridging the gap between creation and consumption still a challenge (metadata, quality, structure, semantics)”
- brauchen eine neue Generation von Daten-Experten

ESFRI Cluster Sorgen/Nöte

identity & integrity

finding, access & re-use

DM & DC

	CRISP	ENVRI	DASISH	BioMed
Data identity				
Data identity continuum				
Software identity				
Concept identity				
User identity management				
Common data standards and formats				
Service discovery				
Service market places				
Integrated data access and discovery				
Semantic annotations and bridging				
Data storage facilities				
Data curation				
Dynamic data management				
Privacy and security				
User Community Forum				
Reference models				
Education & training				

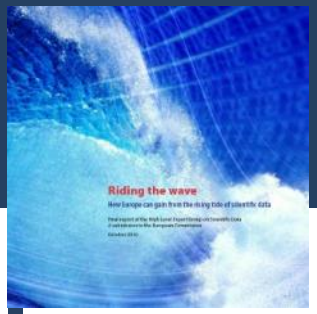
wo geht es lang?
wo könnte es lang gehen!

konkrete Schritte der EC (Oct. 201

N. Kroes EC:

European attempts to build a common data infrastructure

OpenAIRE
and others as well

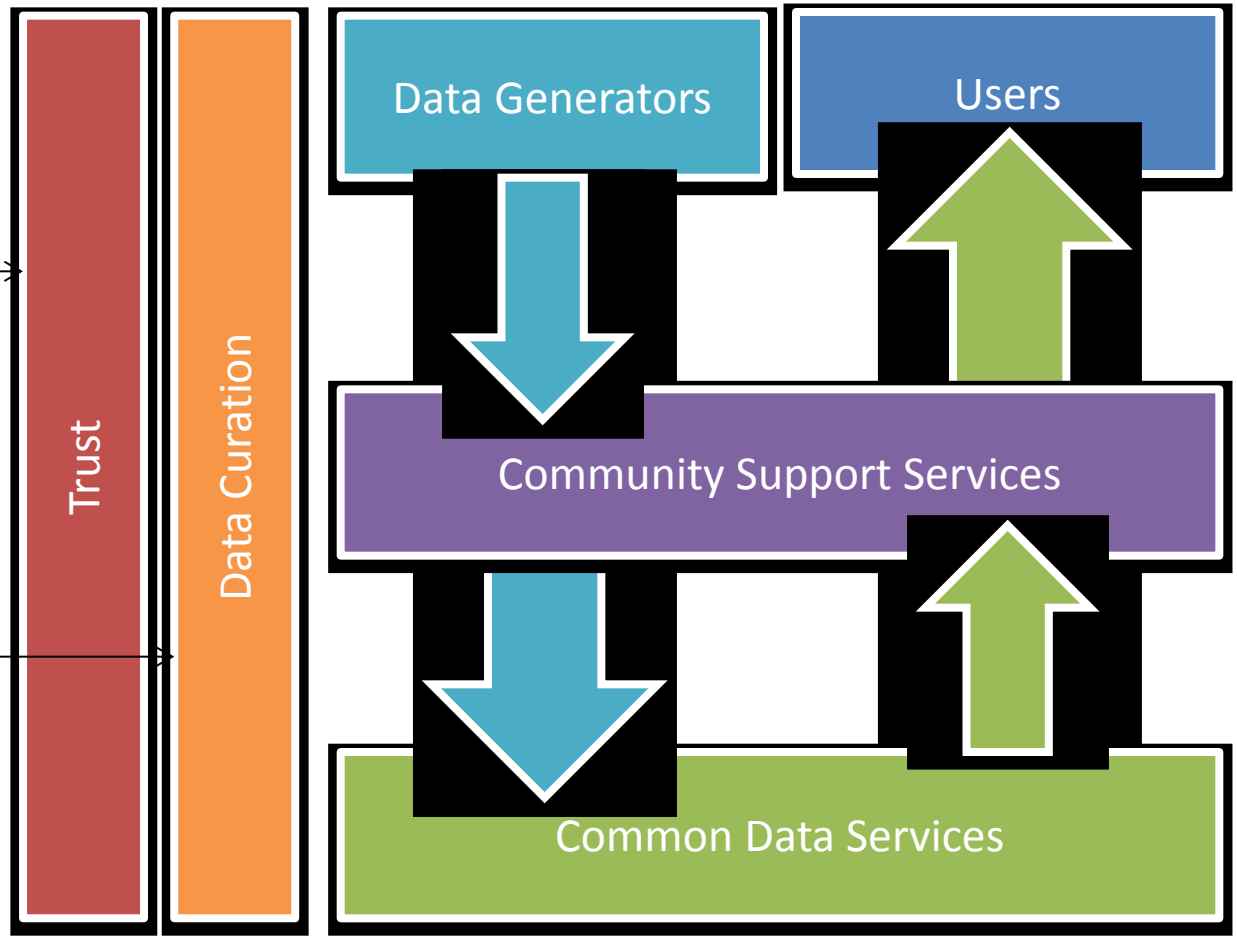


Collaborative Data Infrastructur

MAX-PLANCK-GESELLSCHAFT

organisatorische und kulturelle Nähe sind primär für Vertrauen

Kuration ist eine beiderseitige Aufgabe



EUDAT – Föderation diverser Zentr

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

KIT Karlsruhe Institute of Technology

INGV

CERN

epcc

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

DKRZ DEUTSCHES KLIMARECHENZENTRUM

CINES Centre Informatique National de l'Enseignement Supérieur

CINECA

CSC

RedIRIS

Max-Planck-Institut für Meteorologie

rzg RECHEN-ZENTRUM GARCHING

JÜLICH FORSCHUNGSZENTRUM

Science & Technology Facilities Council

maatG

CERFACS

umweltbundesamt ENVIRONMENT AGENCY AUSTRIA

POZNAN SUPERCOMPUTING & NETWORKING CENTER

SNIC

Trust-IT Services Ltd
Communicating ICT to markets

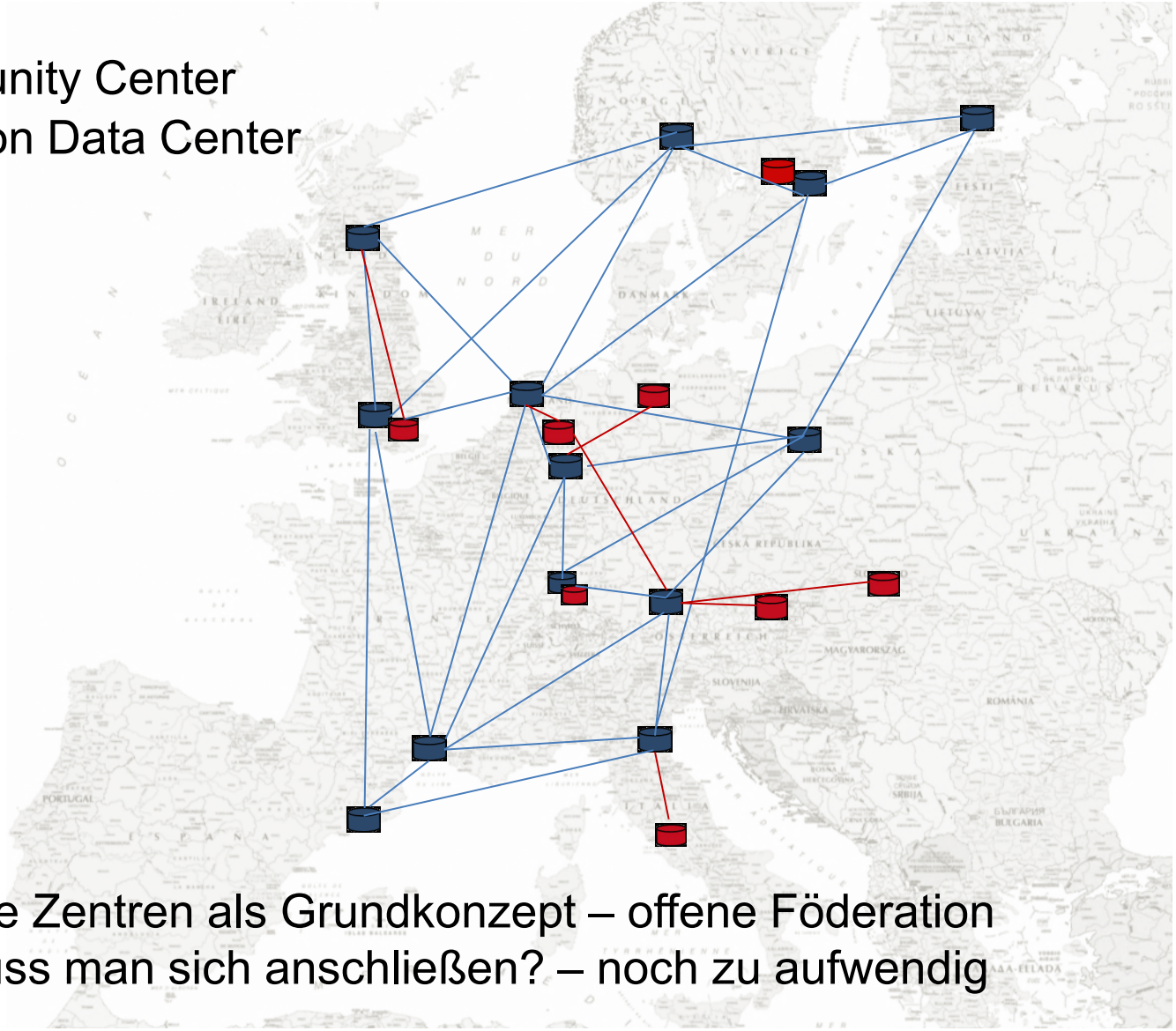
UNINETT sigma

SURF SARA

UCL

EUDAT Knoten (ab 2014 mehr)

- Community Center
- Common Data Center



verteilte Zentren als Grundkonzept – offene Föderation
wie muss man sich anschließen? – noch zu aufwendig

being offered
in progress
to come

EUDAT Service Übersicht

MAX-PLANCK-GESELLSCHAFT

B2FIND
Aggregated EUDAT metadata domain.
Data inventory

B2SAFE
Data curation and access optimization

B2STAGE
Dynamic replication to HPC workspace for processing

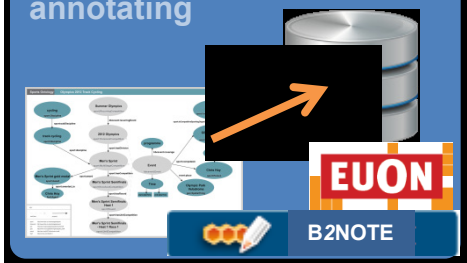
B2SHARE
Researcher data store (simple upload, share and access)

AAI
Network of trust among authentication and authorization actors

PID
Identity
Integrity
Authenticity
Locations

EUDAT Box
dropbox-like service
easy sharing
local synching

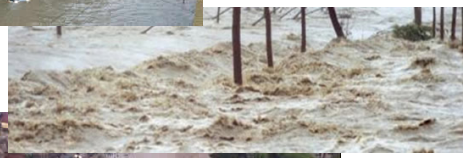
Semantic Anno
checking , referencing and annotating



Generic Workflow
automating data processing

Dynamic Data
immediate handling

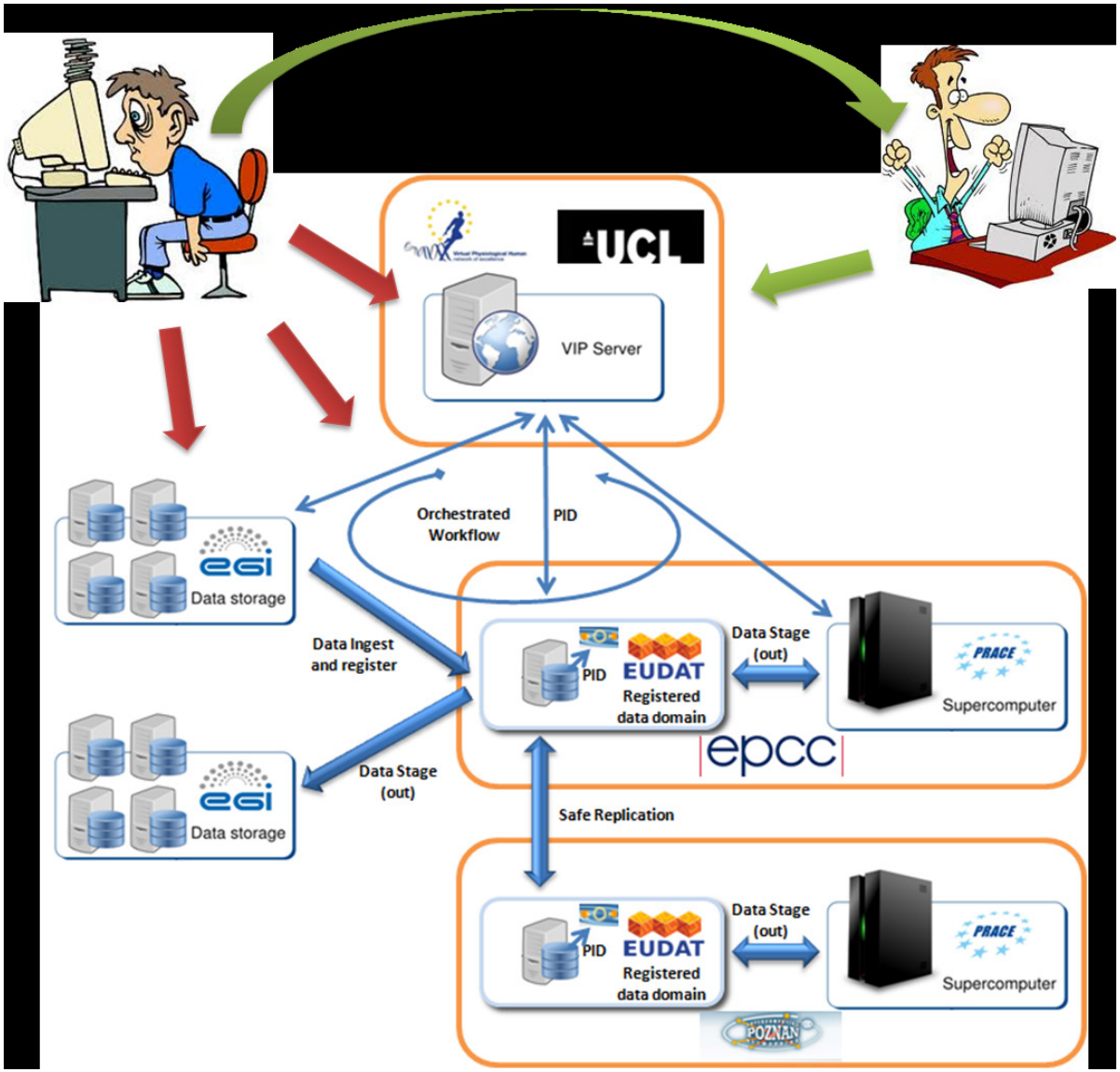
DRIHM – EUDAT für Citizen Data



Community Domain Specific Metadata

Citizens

VPH Replikation + Verarbeitung Big



Replikation auf physischem Niveau ist einfach (File, Cloud-Objects, etc)

Replikation inkl. logischer Information ist komplex (alle machen es unterschiedlich)

Staging zu HPC auch nicht einfach

wo ist das Problem?

- offensichtlich ist alles im Fluss oder?
- außerdem haben wir Infrastrukturen von Google, Amazon etc.
  - viele sind durchaus zufrieden und betreiben Data Mining

- es passt wenig zusammen, d.h. der Aufbau von Föderationen und die Wiederverwendung von Daten ist viel zu aufwendig und teuer
- partielle Reduktion der Komplexität erforderlich
- auf physischem Niveau Vereinfachung durch Cloud APIs
- für Infrastrukturen brauchen wir allgemeine Komponenten und Vereinbarungen
  (AAI, PID, Registraturen, Metadaten, Rechte, etc.)

wir brauchen Vereinbarungen!

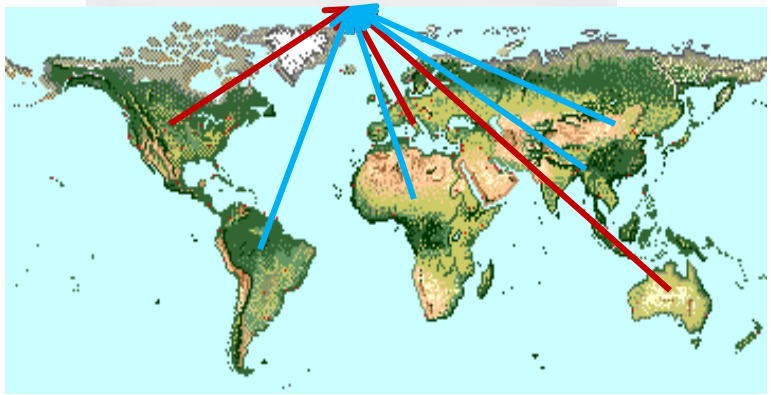
die Geburtsstunde der Research Data Alliance (RDA)

konkrete Schritte der EC Okt 201

N. Kroes EC:

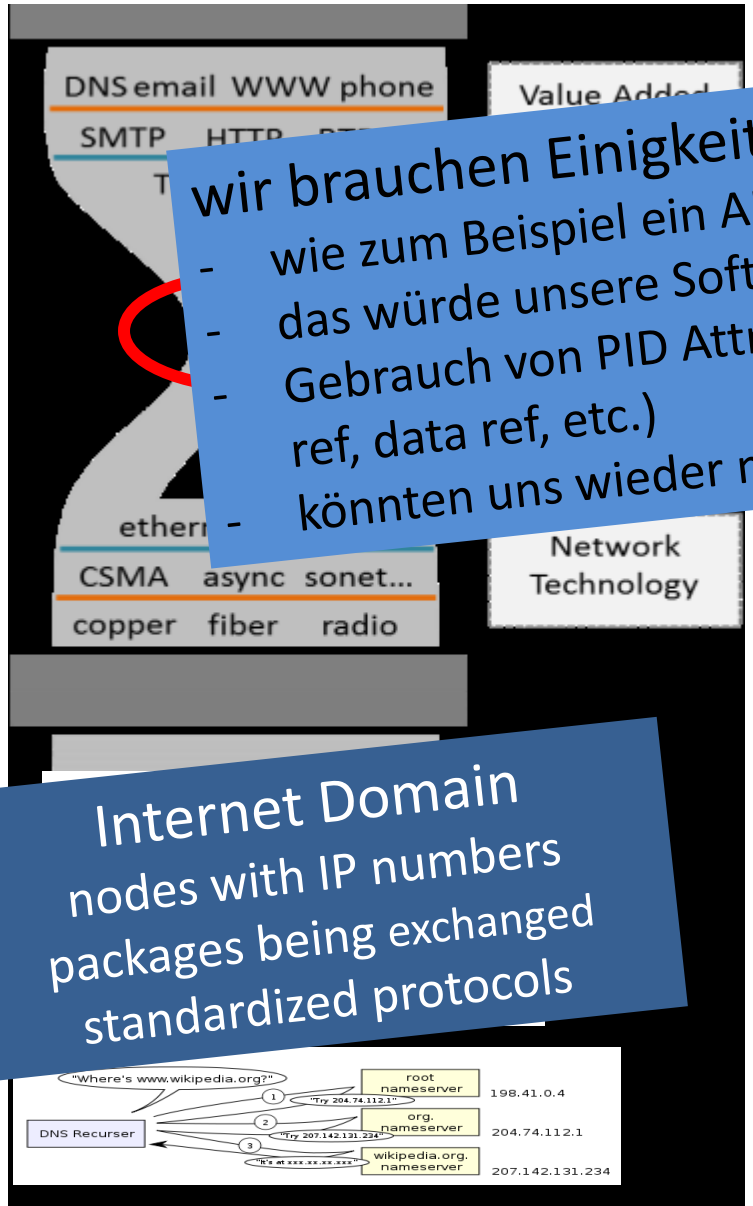
European attempts to build a common data infrastructure

Global attempts to improve data sharing and interoperability



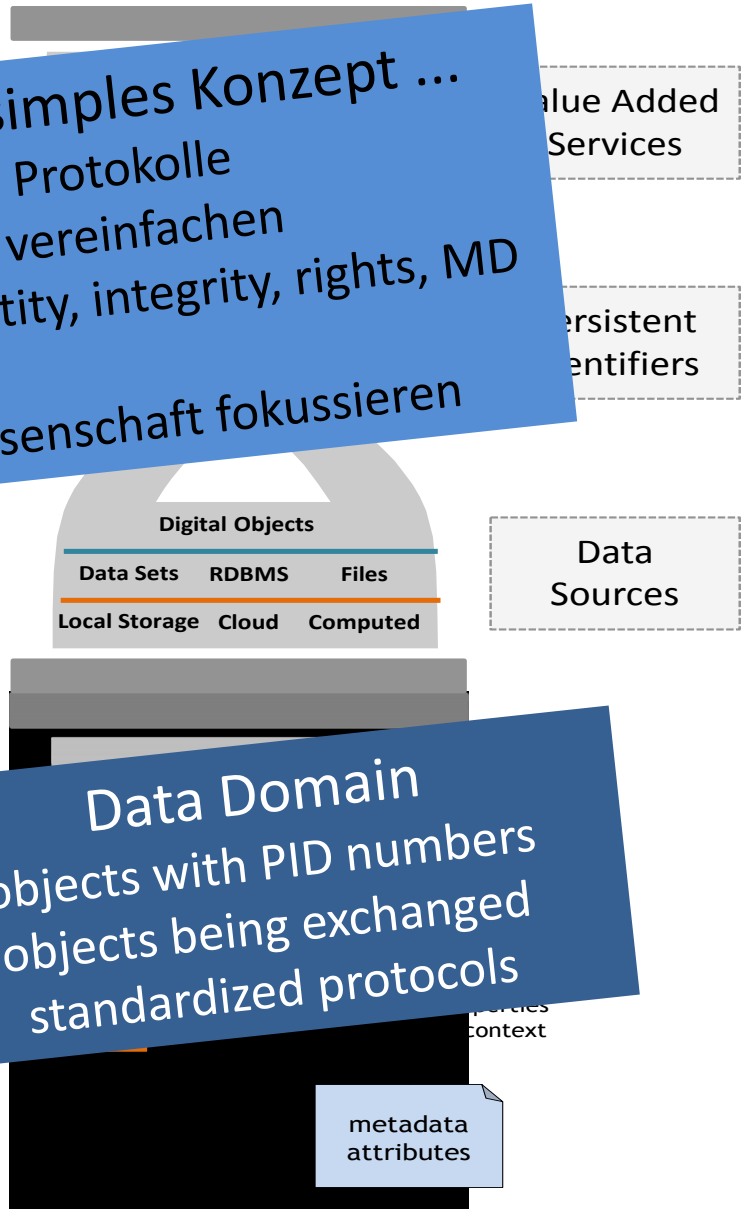
currently supported by NSF, EC and AU more to come soon

Lernen vom Internet



wir brauchen Einigkeit über ein simples Konzept ...

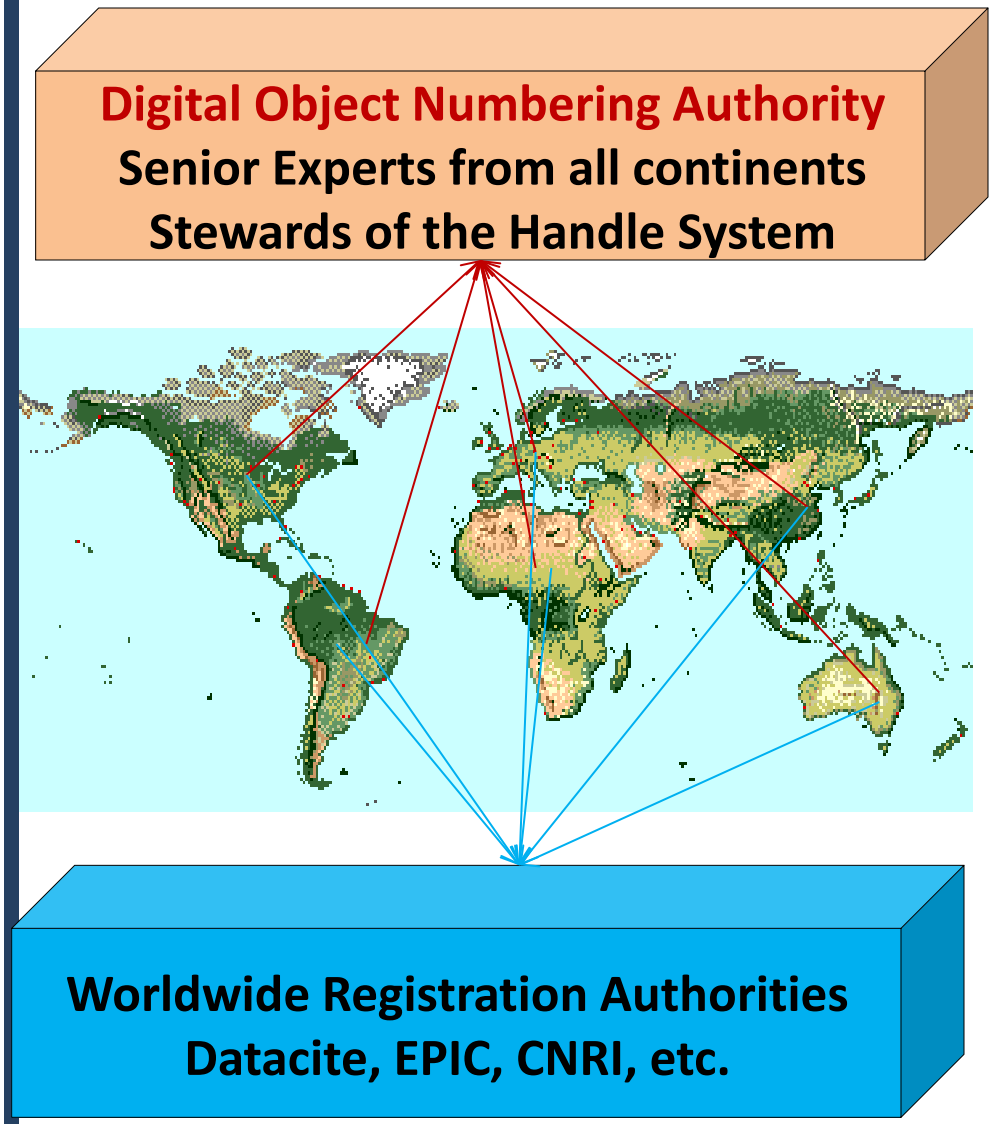
- wie zum Beispiel ein API und einige Protokolle
- das würde unsere Software enorm vereinfachen
- Gebrauch von PID Attributen (identity, integrity, rights, MD ref, data ref, etc.)
- könnten uns wieder mehr auf Wissenschaft fokussieren



Internet Domain
nodes with IP numbers
packages being exchanged
standardized protocols

Data Domain
objects with PID numbers
objects being exchanged
standardized protocols

DONA ist bereits fertig



DONA ist als Schweizer Stiftung unter dem Hut der ITU installiert.

Es wird geleitet von einem international besetzten Board, d.h. das Handle System wird unabhängig von CNRI weitergeführt.

IDF/DataCite, EPIC, CrossRef etc sind Teil des weltweiten und redundatent Service Netzes.

RDA Groups
(here: from the 2nd Plenary in Washington DC , Sept

- **Birds-of-a-Feather**
  - Linked Data
  - Chemical Safety Data
  - Education and Skills Development in Data Intensive Science
  - Libraries and Research Data
  - Cloud Computing and Data Analysis Training for the Developing World
- **Working Groups**
  - Data Type Registries
  - Metadata Standards
  - Practical Policy
  - Persistent Identifier Types
  - Data Foundations and Terminology
  - Data Categories and Codes
- **Interest Groups**
  - Agricultural Data
  - Big Data Analytics
  - Data Brokering
  - Certification of Trusted Repositories (joint with ICSU-WDS)
  - Long tail of Research Data
  - Marine Data Harmonization
  - Community Capability Model
  - Data Publishing (joint with ICSU-WDS)
  - Toxicogenomics Interoperability
  - Research Data Provenance
  - Data Citation
  - Metadata
  - Economic Models and Infrastructure for Federated Materials Data Management
  - Engagement
  - Preservation e-Infrastructure
  - Legal Interoperability (joint with CODATA)
  - Global Registry of Trusted Data Repositories and Services
  - Digital Practices in History and Ethnography

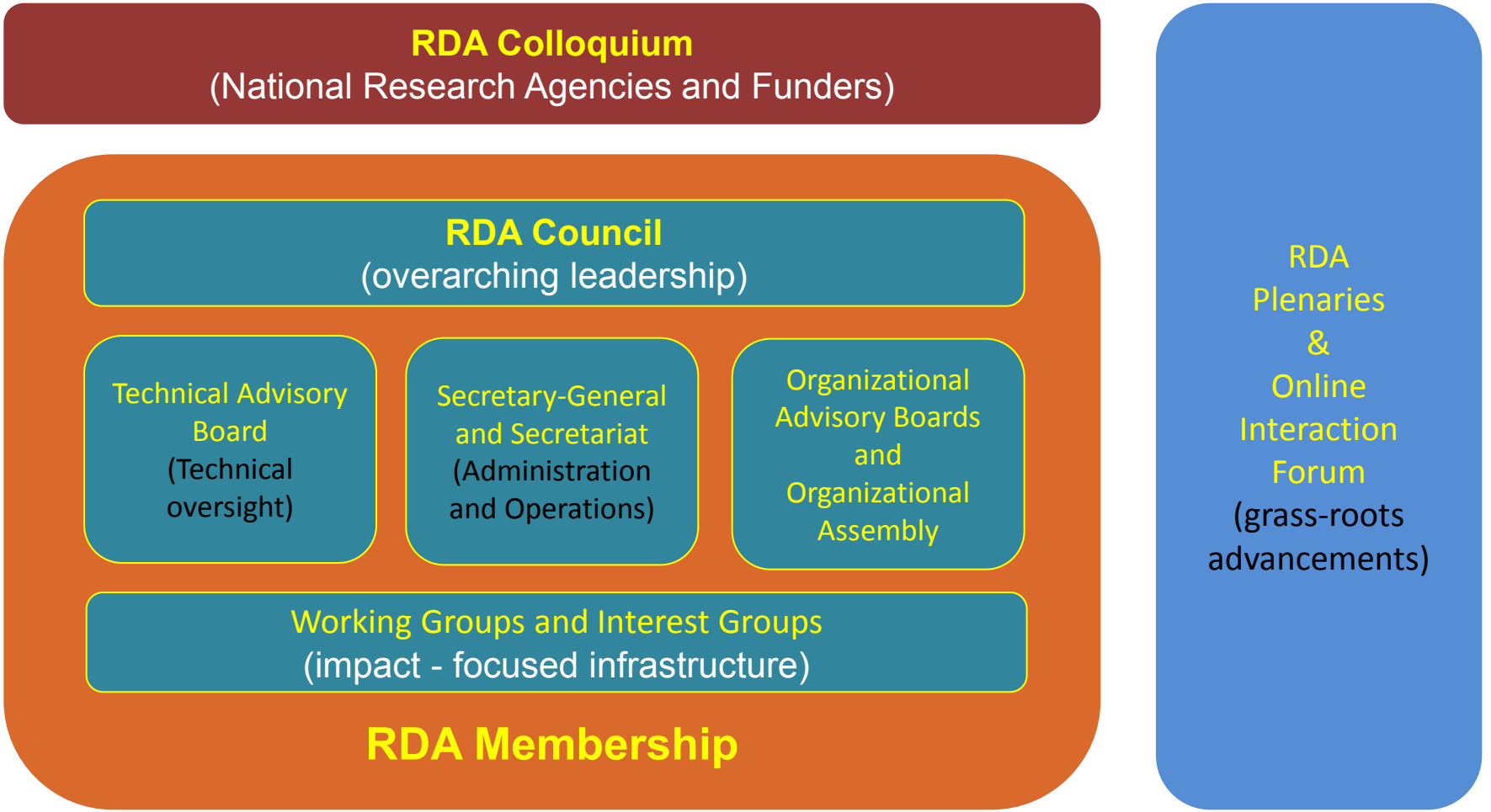
RDA WG Deliverables

- Adopted code, policy, infrastructure, standards, or best practices that enable data sharing
- "Harvestable" efforts for which a 12-18 month effort can eliminate a roadblock
  - Efforts that have a substantial impact within the data community, but might not apply to all
  - Efforts for which scientists and researchers can start today.

RDA Principles
Openness
Consensus
Balance
Harmonization
Community Driven
Non-Profit

inline with G8+O6

RDA Governance



RDA – was ist es also?

- am besten mit dem Internet (IETF) vergleichen
- es ist eine bottom-up Organisation in der “data practitioners” zusammenarbeiten um Daten-Management, -Zugang, -Austausch, -Bewahrung durch das Überwinden von Barrieren viel effizienter machen
- es ist cross-disziplinär angelegt und agiert global, da auch die Wissenschaft global organisiert ist
- natürlich bedarf es einer top-down guidance um alles in Balance zu halten
- das Herz von RDA sind Arbeitsgruppen mit ganz konkreten Zielsetzungen zur Überwindung von Barrieren und Interessengruppen, die in Richtung auf die Bildung von AGs wirken

RDA Plenaries

- Plenary 1: March 18-20, 2013
  - at Gothenburg, Sweden
- Plenary 2: September 16 - 18, 2013
  - Washington, DC, USA
- Plenary 3: Dublin, Ireland
  March 26-28 in 2014
- Plenary 4: Amsterdam, NL
  September 22-24 in 2014

viele andere Meetings zwischendurch

Haben IT Zentren eine Aufgabe?

Zentren in CLARIN-D und DARIAH-D

Aufgaben und Rollen neben den sehr wichtigen Community Zentren:

- Resource provisioning: VMs & Storage:
- Service hosting:
- AAI (shibboleth):
- PID service:
- iRODS Federation (federated storage):
- Monitoring:
- Technical Support (computing centres):
- Operational Security:
- Clarin Center Registry:
- Clarin Workspaces (OwnCloud):

Rolle der IT Zentren

- Infrastrukturen werden durch Föderationen stabiler Zentren mit verschiedenen Services realisiert
- die Datenlandschaft ist komplex und wird auch mehrschichtig bleiben – keiner möchte Monopole und Nähe schafft Vertrauen
- daher wird es Zentren mit community-nahen und solchen mit allgemeineren Funktionen geben (FO, National, EU)
- Zentren brauchen Experten, die pro-aktiv Aufgaben in Zusammenarbeit mit den Wissenschaftlern wahrnehmen
- Zentren brauchen Experten, die bezüglich Föderations-Komponenten, Standards, APIs, Protokolle top-fit sind
- Management, Kuration, LZA, etc. bleiben eine Aufgabe von Zentren mit langfristiger finanzieller Absicherung
- Zentren müssen ihre Policies offen darlegen und sich regelmäßig zertifizieren lassen

Vielen Dank für Ihre Aufmerksamkeit!

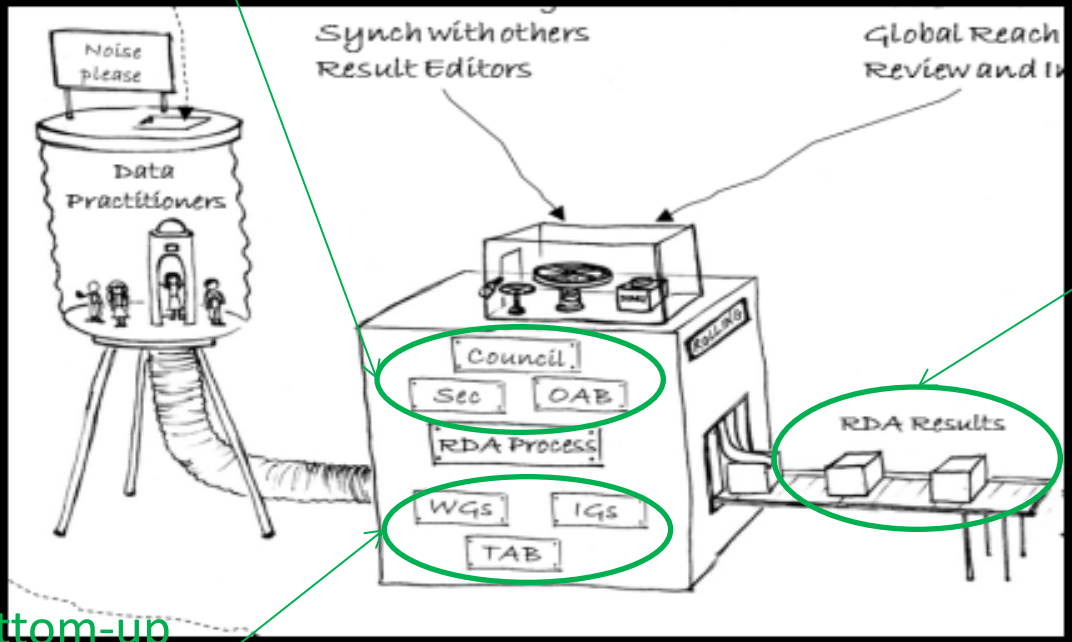
http://www.eudat.eu

http://europe.rd-alliance.org

http://www.rd-alliance.org

RDA Maschinerie

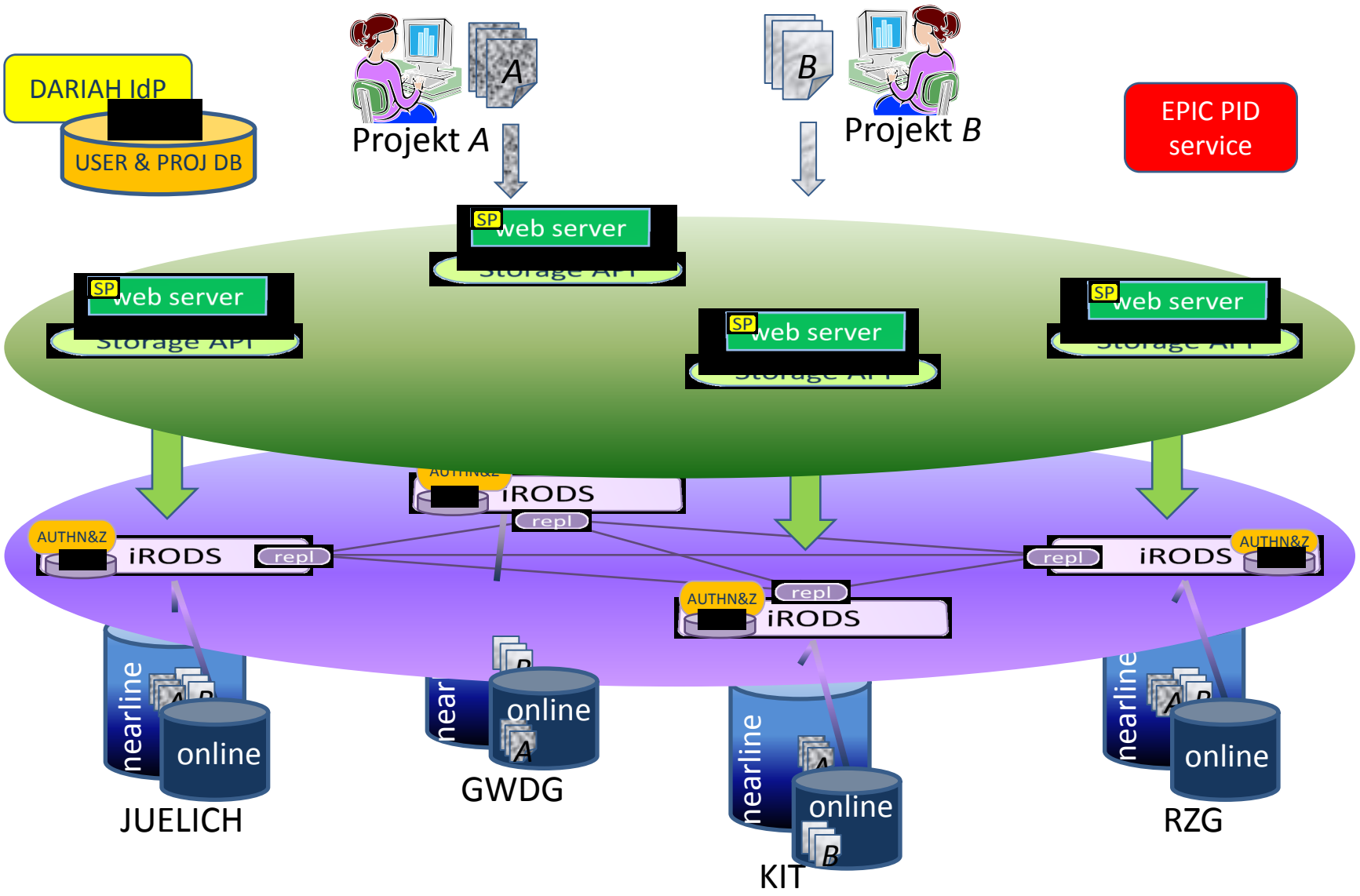
top-down process



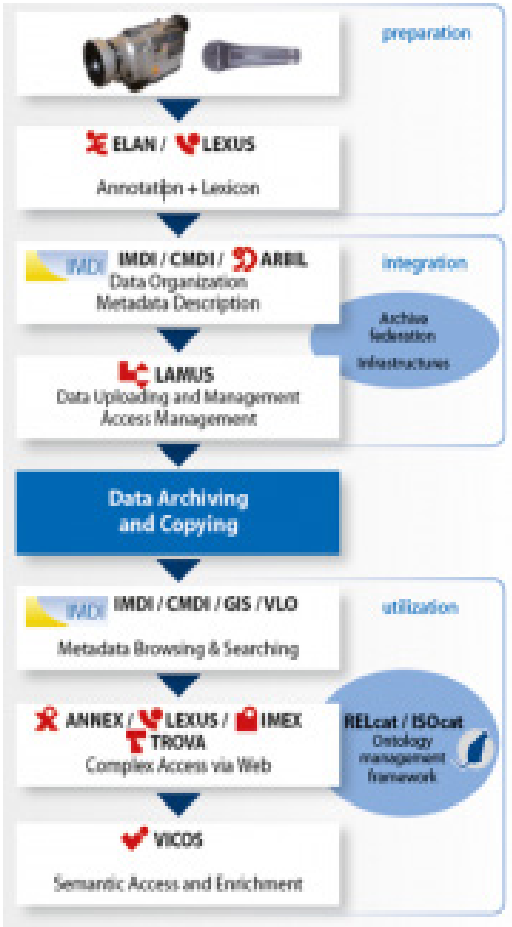
uptake to come

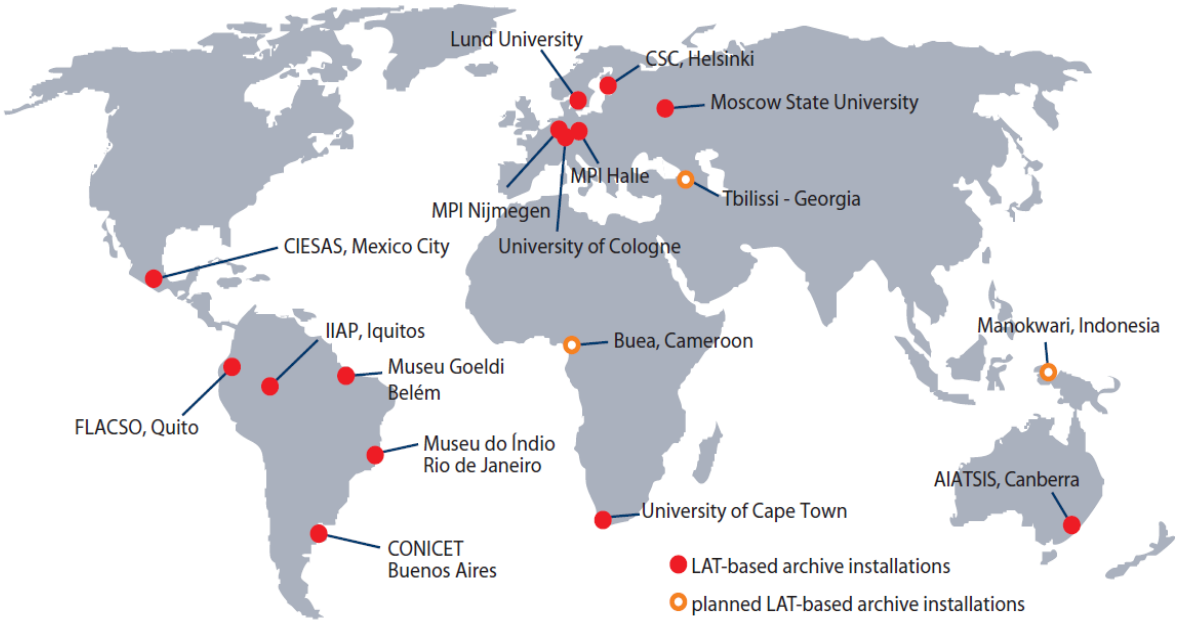
bottom-up process

Community-basierte Infrastrukturen



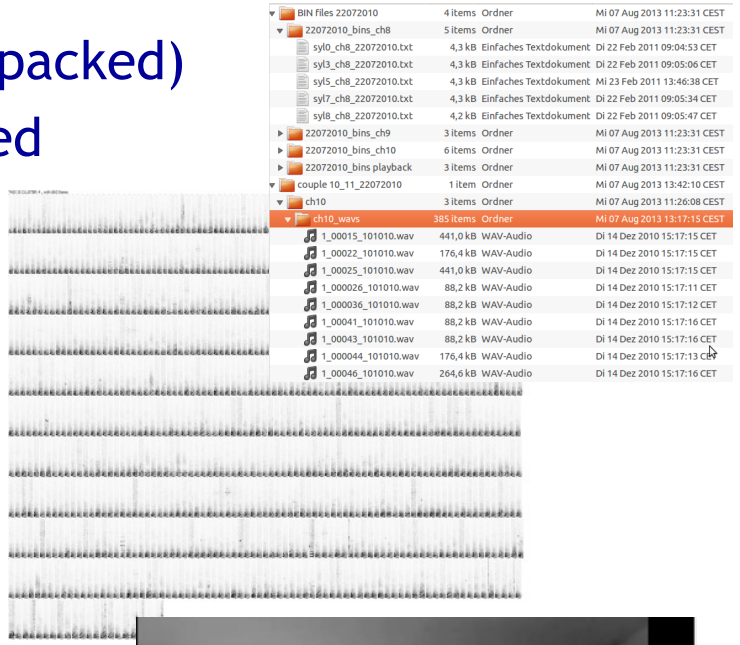
Tool Suite und Archiv-Föderation am M



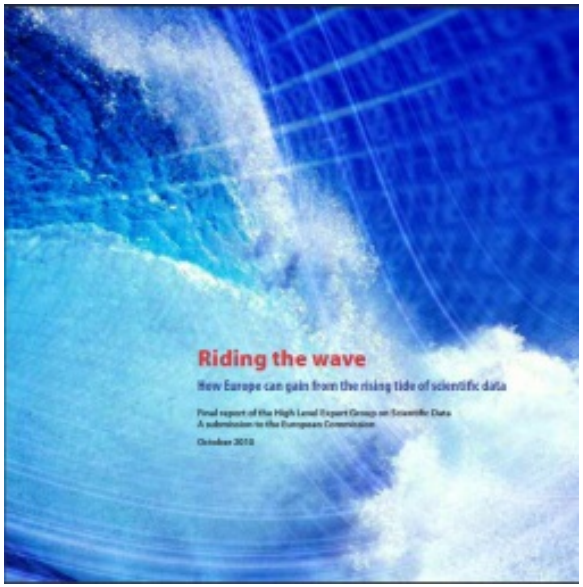


MPI für Ornithologie

- Data volume: 400 TB
- Large number of files: 70 Mio files (to be packed)
- Organized in per experiment folders shared
- by experimentators
- Unstructured heterogeneous data
  - Audio
  - Videos
  - Images
  - Binary files (aggregated datastreams)
  - Text files
- Metadata available in various forms,
  in many cases no structured **digital** metadata available.



einige Kern-Aussagen

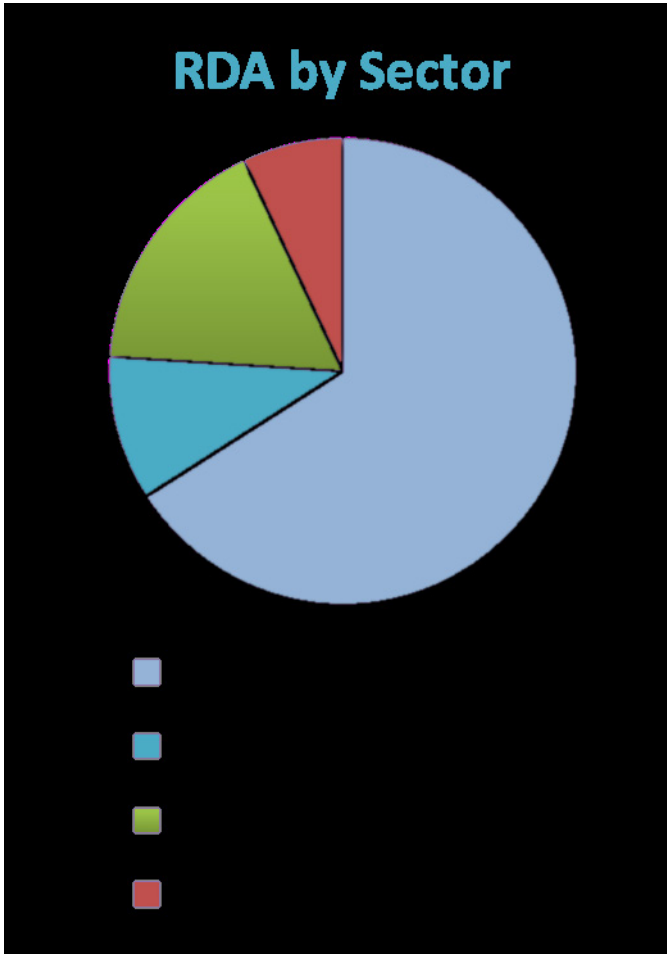


Riding The Wave (EC's HLEG on Scientific Data)

"The emerging infrastructure for scientific data must be flexible but reliable, secure yet open, local and global, affordable yet high-performance. Obviously, this is a tall order – and there is no one technology that we know today or can imagine tomorrow to achieve it all. Thus, what is needed is a broad, conceptual framework for how different companies, institutes, universities, governments and individuals would interact with the system – what types of data, privileges, authentication or performance metrics should be planned. This framework would ensure the trustworthiness of data."

RDA Community Current Status: ~1,300 participants from 50+ countries

1. Albania
2. Australia
3. Austria
4. Bangladesh
5. Belgium
6. Bolivia
7. Botswana
8. Brazil
9. Bulgaria
10. Canada
11. China
12. DR Congo
13. Costa Rica
14. Czech Republic
15. Denmark
16. Estonia
17. Finland
18. France
19. Germany
20. Greece
21. Iceland
22. India
23. Iran
24. Ireland
25. Italy
26. Japan
27. Kyrgyzstan
28. Kuwait
29. Mexico
30. Netherlands
31. New Zealand
32. Norway
33. Palestine
34. Poland
35. Portugal
36. Russian Federation
37. Rwanda
38. Serbia
39. Singapore
40. Slovenia
41. South Africa
42. South Korea
43. Spain
44. Sweden
45. Switzerland
46. Taiwan
47. Turkey
48. United Arab Emirates
49. United Kingdom
50. United States
51. Vatican City
52. Venezuela

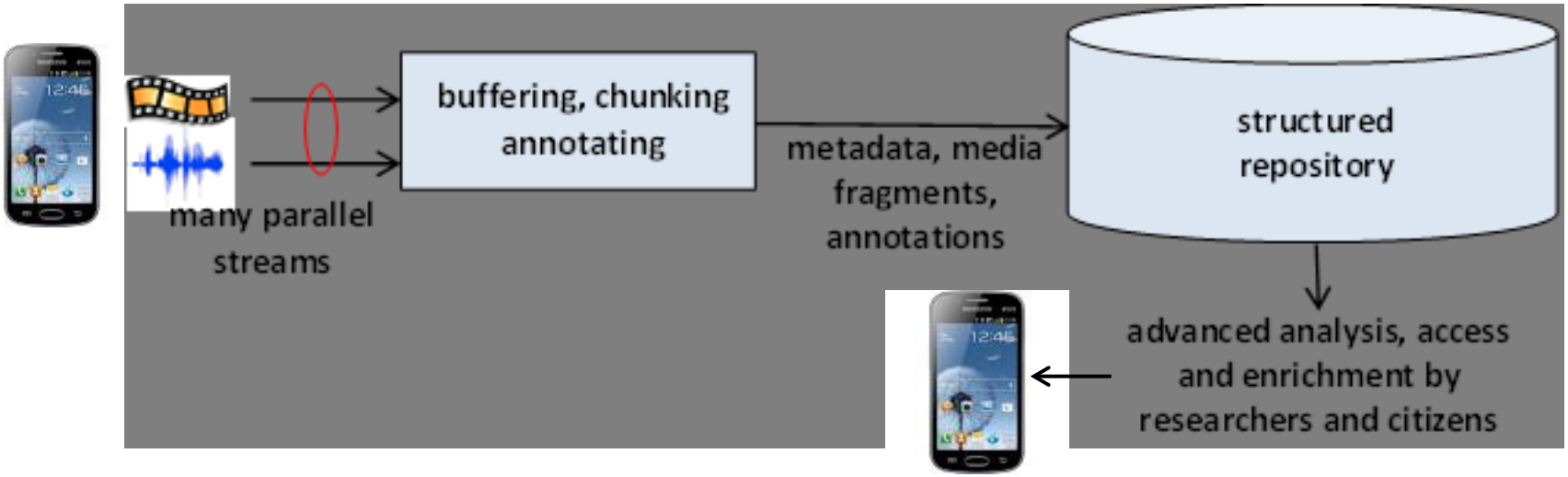


answers to questions

- *RDA is NOT an infrastructure – but a machinery to quickly push agreements (specifications, running code, etc.)*
- *RDA is open to everyone dealing with data and signing the principles – it is a neutral meeting place also for RI and eInfra experts to remove barriers*
- *RDA is very much inspired by some RI and eInfra –inspiration by concrete problems is essential*
- *the user community should be global and cross-disciplinary*

  yet not so evident how to bridge all initiatives
- *RDA users should are data practitioners within RI & eInfra*
- *Interoperability is essential to reduce costs for working with data and thus democratize science and enable innovation*
- *all RDA is focusing on reducing barriers for dealing with data*

massives Crowd Sourcing im MPI



- crowd sourcing schon im Einsatz - noch zu viel Amateurismus
- massives CS im Kommen – viele VP und MD mit Sensoren
- 10 min * 100 P/Tag multimedia Aufnahmen (H.264) = 100 GB/T
- benötigen Maschinerie zur Reduktion/Annotation, zum DM und für das Feedback zu Teimnehmern
- alles muss hochgradig automatisiert sein